This week：

1. configuration：

   It takes me two days to install Scala and Spark on computer. I use Intellij-IDEA and reinstall java SDK 1.8. However, cluster of our lab needs to be repaired, so I do not test our program on Spark.

2. Project (data program)：

I use java derives one-day trajectory data for a total of 36GB and then derives one hour of the data as the test data.

3. Paper write：

I write Introduction of the paper, and I will continue this part of work next week:

With the advances in Sensing technologies and large-scale computing infrastructures, a variety of trajectory data have been produced.

Trajectory data records the location (e.g., latitude and longitude) and corresponding time information of the moving objects.

In addition to spatio-temporal attributes, other types attributes such as speed or direction of moving objects are also recorded.

Many fields such as urban planning, crowd mobility analysis and traffic management rely on trajectory data to efficiently enrich their knowledge and assist in analysis.

Combination of Trajectory and its semantic information is a crucial step for trajectory data analysis~\cite{Alvares2007A}. Several researches focus on integration of semantic and trajectory~\cite{Alvares2007A}, retrieval~\cite{Aldohuki2016SemanticTraj,Hu2016Top}, management~\cite{Richter2012Semantic} and data mining for semantic trajectories.

However, because of the limitations of the sampling rate, many trajectories have low-resolution and low-precision, e.g., Mobile phone location data records the location of a mobile phone every 20 minutes, each sampling point is based on the base station to record, The location accuracy is 100- to 2000- meters depending on the cell tower coverage in the area which deliver a lot of problems about the trajectory query, analysis and exploration. For example:

\begin{itemize}

\item \textbf{Scenario 1:}

Analyst want to query the people who visit ZhongShan park and the people who go to Children's Hospital which next to ZhongShan park. Because of low accuracy of the mobile phone trajectory

data, it is a difficult work to distinguish these two group of people during to the high-spatio-similarity of ZhongShan park and Children's Hospital.
\end{itemize}

%在传统的轨迹系统中，用户想要找到这两类人群的方法：首先通过 POI 的名称找到地点位置，然后通过地理查询检索经过的轨迹。由于地理采样的不精确性，需要给地理检索条件加上一个与目的地相对应的概率参数来约束查询结果，最简单的例子就是轨迹点距离 POI 位置越近，那么他经过该 POI 的概率就越高。

%POIs 的赋值，

%不平衡的余弦相似性分析，主要解决的问题是查询的问题，我们有很多相关的关键词，而这些关键词可以通过 POI 类别的统计值来计算一下查询出数据和输入条件之间的余弦值差异，从而定义查询结果的匹配程度。由于 POI 的种类，或者 POI 名称比较多，而在实际的应用中，用户可能只关心某一类 POI，所以我们可以采用偏差的余弦相似度分析来过滤掉无关的属性。

%关于数据属性的描述，或者说关于数据的数据称为元数据。元数据可以帮助程序完成对于数据的分类，分层，聚类等信息，完成有监督的聚类。

%概念分层的部分：POI 的类别已经分好了层。

In summary, our contributions is a novel retrieval method for low sampling rate and Low accuracy trajectory data and a visual analysis system to analyze uncertainty trajectories which overcome the bottleneck for querying and understanding the low-resolution low-precision trajectory data.
First, we describe a new data model that utilize POIs data and experts knowledge to assign low resolution trajectories with semantic information.
Second, to retrieval trajectories, we uses textual index and rank engine to manage such trajectory data and achieve fast query efficiency.
Finally, we proposed a visual analysis system that supports the querying, visualization and exploring of semantic trajectory data for various analysis tasks.
Several usage scenarios are proposed to show the usability of our system.

The remainder of the paper is organized as follows: In section 2, we introduce the related work cover semantic trajectory query, management and visualization. In section 3, we present our data

model, textual index and rank engine which achieve fast query efficiency. We proposed our visual analysis system in section 4 as well as several usage scenarios in section 5. Finally, in section 6, we discuss our method and give some future work.

4. Paper reading：

   I read the book <data mining>. About 4 chapters. Many work like data clean, preprocessing, Calculation of similarity and probability are useful, and I read the related references:

   - 计算数据之间相似度的算法（都是一些比较老的文章）：

     可以用相异性矩阵衡量对象之间的临近程度。其中有对称的二元相异性，非对称的二元相异性，标称属性的临近性等等。 [5, 6, 7, 8]

     我们有很多相关的关键词，而这些关键词可以通过 POI 类别的统计值来计算一下查询出数据和输入条件之间的余弦值差异，从而定义查询结果的匹配程度。可以使用卡方检验对数据属性进行相关性的分析。

   - 数据清洗中需要处理的东西：

     数据的缺失，噪声，数据规约和数据变换都是我们的工作前期需要考虑的东西，本周完成了数据规约工作。

读了这部分的书对我的工作有了一些启发。相似性的计算算法有了最初步的计算。可以先进行测试一下，看看有没有结果。

参考文献：

1. Alvares L O, Bogorny V, Kuijpers B, et al. A model for enriching trajectories with semantic geographical information[C]// ACM International Symposium on Advances in Geographic Information Systems. ACM, 2007:22.
2. Al-Dohuki S, Kamw F, Zhao Y, et al. SemanticTraj: A New Approach to Interacting with Massive Taxi Trajectories.[J]. IEEE Transactions on Visualization & Computer Graphics, 2017:1-1.
3. Hu H, Li G, Bao Z, et al. Top-k spatio-textual similarity join[C]// IEEE, International Conference on Data Engineering. IEEE, 2016:1576-1577.
4. Richter K F, Schmid F, Laube P. Semantic trajectory compression: Representing urban movement in a nutshell[J]. Journal of Spatial Information Science, 2012, 4(4):3-30.
5. Hartigan J A. Clustering algorithms[M]. Wiley, 1975.
6. Jain A K, Dubes R C. Algorithms for clustering data[J]. Technometrics, 1988, 32(32):227-229.
7. Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis[M]// Biometrics. 1990.
8. Arabie P, Hubert L J, Soete G D. Clustering and Classification[J]. Biometrics, 1996, 3(53):109-128.